

时间序列的检索

1、时间序列相似性问题

“Using Multi-Scale Histograms to Answer Pattern Existence and Shape Match Queries over Time Series Data”

是 2005 年的论文，作者 Lei Chen 目前在香港科大，他总结基于相似性的时间序列检索主要有两种方式：pattern existence queries 和 shape match queries。

Pattern existence queries 用户关心时序数据中 general pattern，而不注重细节，例如“Give me all the temperature data of patients in last 24 hours that have two peaks”。其中噪声、振幅、时间平移和时间比例等因素需要考虑。一些近似方法将时间序列转换为字符串，应用字符串匹配技术来发现模式。

Shape Match Queries 用户关心与查询数据相似的 movement 形状，如“Give me all stock data of last month that is similar to IBM's stock data of last month”。通常应用距离函数如欧式距离、DTW（Dynamic time warping）和 LCC（Low-complexity chase）计算两个时间序列。

但是在很多情况下，可能同时需要按照两种情况检索，例如用户一开始对所有具有某种特定模式的时间序列感兴趣，可以快速使用 pattern existence queries。随后用户需要用 shape match queries 对查询的结果进一步搜索，获取与结果中某一类时间序列相似的序列。

为此作者提出了能够同时满足两种检索的方法：多尺度直方图。

所谓时间序列直方图其实是对时间序列的每个值做规范化操作，构成的阶跃形式。

对于时间序列 R 可形式化表示为 $R = [(r_1, t_1), \dots, (r_N, t_N)]$ ，对于时间序列集合 $D = \{R_1, R_2, \dots, R_L\}$ ，每个时间序列 R_i 可以以均值和方差规范化表示为：

$$Norm(R) = [(t_1, \frac{r_1 - \mu}{\sigma}), \dots, (t_N, \frac{r_N - \mu}{\sigma})]$$

文中使用 weighted Euclidean distance（WED）来计算两个直方图的相似度。

时间序列直方图能够表达时间序列的整体结构，但是在检索时，往往只需要对局部结构搜索，因此文中继续提出多尺度的时间序列直方图，将整个时间序列分层划分为 $2^{\delta-1}$ 层。每一层按 2 的指数倍分割， $\delta = 1$ 表示整个时间序列，为 2 时将时间序列划分为两段，依次类推...

另外有两篇论文提出的搜索方法是目前在单变量时间序列检索中最好的

[1] J. J. Shieh and E. Keogh. ISAX: Disk-aware Mining and Indexing of Massive Time Series Datasets. Data Min. Knowl. Discov., 19(1):24–57, 2009.

[2] C. Traina, R. Filho, A. Traina, M. Vieira, and C. Faloutsos. The Omni Family of All-purpose Access Methods: A Simple and Effective Way to Make Similarity Search More Efficient. The VLDB Journal, 16:483–505, 2007.

2、多变量时间序列

“Fast and Flexible Multivariate Time Series Subsequence Search”

是2010年ICDM论文，在以前的论文中关于多变量时间序列(MTS)通常同时对一组固定的变量

进行查询，而无法选择变量。Bhaduri, K.针对航空安全数据，提出了两个多变量查询的例子：

1. Return all the flights (a subset of the MTS) where the altitude monotonically changes from 10000 ft to 5000 ft, speed varies between 300 knots to 200 knots, and landing gear is down. Such combination of parameter values may be precursors to unstable approaches while landing.

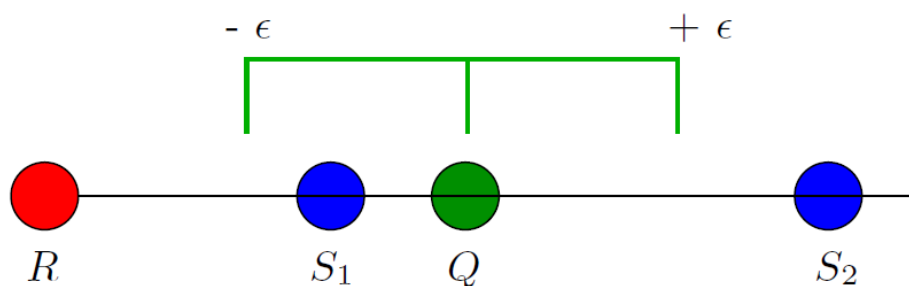
2. Return all the flights where the aircraft is climbing at 100 ft/s with flaps not withdrawn. There may be a time delay between these two sequences.

在这两个问题中，时间序列的查询条件由多个变量组成。在MTS检索中用户可能随时选择变量对任意时间点进行查询，用户选择覆盖的时间范围是任意长度的，不同时间序列的波动可能存在时间偏移，用户可能对查询结果允许的最大误差有要求（阈值）。MTS搜索算法必须要返回所有匹配的结果，同时满足时间偏移和阈值的约束。

本文作者根据现有的单变量时间序列算法，提出了基于列表（LBS）和基于R树的搜索算法。对第 v 个变量的单变量查询，暴力的做法是寻找所有的 k 近邻，就是与所有长度为 L （输入序列长度）的时间序列比较，而且每个偏移都比较一次，这是非常耗时而且不切实际的。典型的数据挖掘方法对其进行加速，具体做法是找到一个距离度量的下边界，用这个边界对不可能的候选进行剪枝。这要求下边界应该是：（1）要比计算所有子序列的代价小；（2）与原始的距离度量要足够靠近，否则无法充分剪枝。在本文中，作者针对 k 近邻的查询无需逐一计算每个下边界，就能够充分剪枝。

下图说明了具体的剪枝过程。首先随机选择 R 的子序列，计算与其余的子序列的距离。然后，根据到 R 的距离对这些子序列排序。 S_1 和 S_2 是最后查询结果。上述两步只是在查询进行前做的而且只需做一次。当查询 q 到达时，只要计算 q 与 R 之间的距离。所有在范围

$[d(q^v, R) - \epsilon, d(q^v, R) + \epsilon]$ 都被剪枝。



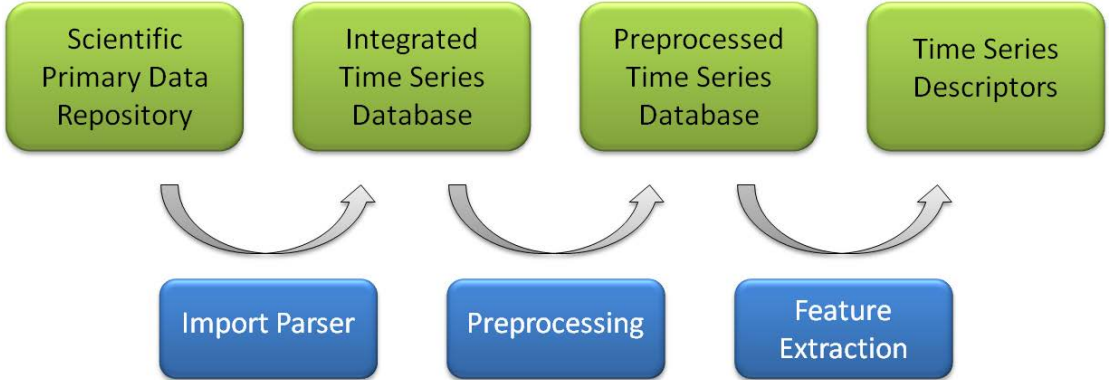
3、时间序列可视化

A Visual Digital Library Approach for Time-Oriented Scientific Primary Data

这是2011年在DL会议上发的论文，是一个前瞻性的工作。工作只是一个雏形，有待完善，但是其思路非常好。撇开关于DL的描述部分，只看可视化检索部分。时间序列的可视化与搜索中最重要的是相似性度量。主要有三种相似性计算方法：基于原始数据、基于模型（比较统计模型）和基于特征（比较特征描述符间的距离）。特征提取是应用较多的方法，易于索引、实现简单而且非常有效。主要的特征提取有傅里叶分析、聚合函数和离散方法。另外在时间序列搜索中，需要区分全局搜索和部分搜索。全局搜索需要比较所有的数据，而部分搜索则需要定义sliding-window或分段方法。基本过程包括四个阶段（如下图）：数据存储、特

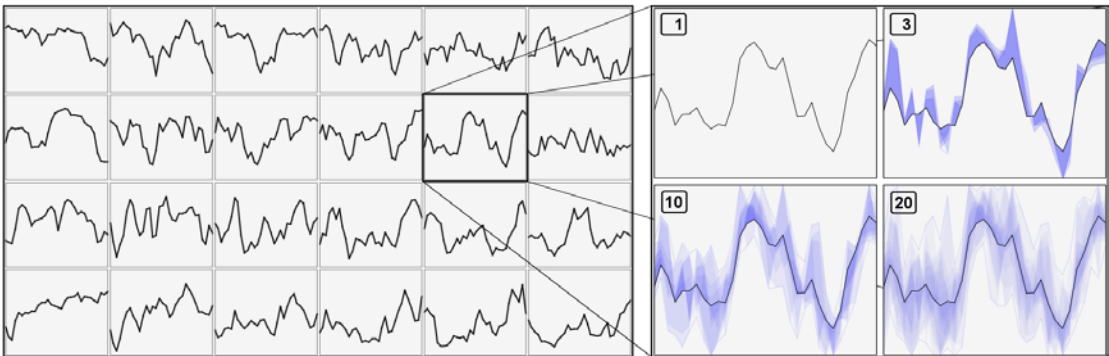
征提取、检索和可视化。这里主要关注特征提取和可视化。

特征提取过程如下图，从中可见特征提取为后面的相似性计算做好了准备。数据预处理主要是执行规范化操作，包括数据离散化、转换、差值以及空值补全、补全空值以及删除离群点等。特征抽取应用了特征描述符抽取方法。虽然傅里叶变换能更好地描述特征，但是本文只用了基于聚集的方法。

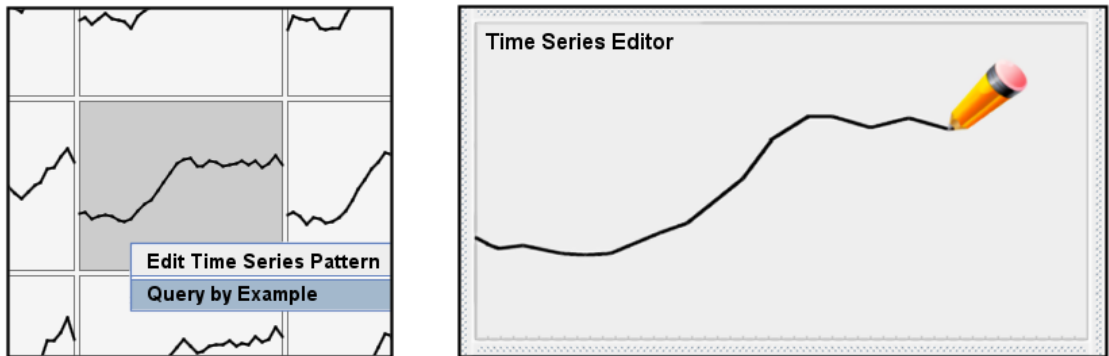


可视化本文是用了visual catalog（如下图）帮助用户开展数据探索，作者认为两方面体现了visual catalog的有效性：

- （1） 反应了不同时间序列的相似性关系
- （2） 减少了显示的数据基数，有利于发现数据中最突出的模式。



所有的可视化结果以SOM进行布局，保证相似的结果能聚集在一起。查询过程中，本文提供了两种交互方式：example-based和sketch-based（见下图）。



同时该系统还支持关键词搜索，对于某个用户下的数据会高亮显示。